Rejoinder: Error in Confidence Judgments

Edgar C. Merkle
Wichita State University

Trisha Van Zandt
Ohio State University

Winston R. Sieck
Klein Associates

Rejoinder: Error in Confidence Judgments

People are sometimes overconfident in their decisions, at least in laboratory settings. Or are they? Erev et al. (1994) provided a demonstration that error could produce an overconfidence effect, depending on how data are analyzed. Juslin et al.'s (2000) position was stronger; they suggested that there was "little or no evidence for an information-processing [overconfidence] bias in human judgment" (p. 388). While we agree with both Erev et al. and Juslin et al. that error can produce overconfidence effects, our paper outlined a problem with Juslin et al.'s technical approach to separating error from "real" overconfidence: such a procedure, when applied to confidence judgment data, will happily remove overconfidence effects regardless of their source (random error or "real" overconfidence due to systematic biases). Furthermore, their procedure gives unrealistically large estimates of the extent of such error. The fact that the procedure can lead to erroneous conclusions means it cannot be used to support such arguments as there being "little or no evidence for an information-processing bias in human judgment."

Olsson et al.'s reply to our paper touches on three issues: 1) our error estimates are implausible and inconsistent with other findings; 2) our demonstrations notwithstanding, measurements of bias can only be taken seriously when error is accounted for; and 3) we misrepresent their position. We discuss each issue in turn below.

Estimating error. There are many potential sources of error in confidence judgment tasks, and we must be careful to specify which we are talking about when arguing for what is or is not reasonable. For example, "response error" is different from "trial-by-trial error."[1] After the decision, which may involve memory search and other cognitive

processes, an internal feeling of confidence may have some error. There may be additional error in mapping this internal confidence to an external, experimenter-defined confidence scale. Like Erev et al. (1994), we define response error to be error that occurs in the mapping from internal confidence to external confidence, and we assume that such error occurs only during this mapping process.

However, Juslin et al.'s (1997) response error model (which is the basis of Juslin et al.'s (2000) analysis) assumes that there is no error in a judge's internal confidence. This means that response error in the model represents everything in the elicitation process (perhaps including systematic biases) that might cause confidence to fluctuate over repeated assessments of the same item. This represents a more general trial-by-trial error, as opposed to the specific response error defined by Erev et al. (1994) and used by us. For example, the trial-by-trial error incorporates retrieval mechanisms, which are typically modeled as stochastic processes and which have been directly linked to overconfidence (Sieck & Yates, 2001). It seems clear that, on this issue at least, our different definitions of response error are the source of the disagreement between us and Olsson et al. (2008). A careful examination and mathematical treatment of error sources in confidence judgment could resolve this disagreement.

With regard to specific estimates of these different sources of error, the 2%-7% figure that we use in our paper is neither response nor trial-by-trial error. Instead, it represents the percent of *responses that are incorrect due to response error*. Olsson et al. (2008) point out (manuscript pp. 6-7) that empirical estimates of *response error variance* are greater than 2-7%. For instance, Budescu et al. (1997) estimate that 20% of the total

variability in confidence can be attributed to response error. Juslin et al. (2003) estimate, via repeated assessments, that: 1) response error represents 17% of the total variability in confidence; and 2) the response error variance is .015. However, because we were talking about percentage of incorrect responses due to response error (instead of percentage of total variability due to response error), the larger estimates noted by Olsson et al. (2008) have little bearing on our estimates.

Furthermore, Juslin et al.'s (2003) 20% estimate obtained by repeated assessments probably does not represent the contribution of response error alone. There are many stochastic components in the confidence elicitation process that can cause trial-by-trial fluctuations in confidence (cf., Merkle & Van Zandt, 2006). As we describe in our paper, a way to further isolate response error would be to study confidence for test items that participants "know." We would expect participants' responses to always be 100% in these situations, unless response error inadvertently causes them to stray from 100%.

Finally, after discussing response error estimates, Olsson et al. (2008) suggest that researchers might focus on reducing unsystematic error in confidence (as opposed to debiasing judges; manuscript p. 7). Averaging across multiple judgments of the same stimulus (e.g., Budescu, Wallsten, & Au, 1997; Juslin et al., 2003) or manipulating response procedures may accomplish this. Of course, as we showed in our paper, it may be difficult to discern whether we are reducing systematic or unsystematic error with such procedures. We agree with Olsson et al. (2008), however, that existing debiasing efforts have room for improvement. Specifically, there is too much focus on overconfidence measures: why should we expect people to have the ability to report

confidence judgments that exactly match proportion correct? Keeping in mind the uses of confidence in applied situations (eyewitness testimony, weather and financial forecasting, etc.), we might instead focus on the extent to which confidence allows us to distinguish correct choices from incorrect choices. Furthermore, most confidence studies use data that have been averaged across participants (e.g., Lichtenstein & Fischhoff, 1977; Suantak, Bolger, & Ferrell, 1996; Windschitl & Chambers, 2004), which may yield misleading results. Clearly specified mathematical models of choice confidence (Erev et al., 1994; Merkle & Van Zandt, 2006; Wallsten, 1996) embedded within hierarchical models (e.g., Rouder & Lu, 2005) would allow us to address these issues directly.

Error can obscure bias. We agree completely with the authors' position that demonstrations of bias become convincing only if error explanations can be discounted (manuscript p. 5). However, there are data that make the error explanation difficult to sustain. Consider Experiment 2 from Sieck, Merkle, and Van Zandt (2007), where participants' confidence changes when they are required to explain why each alternative may be true. Consider also Fischer and Budescu (2005), where participants viewed different types of triangles and assigned them to one of three categories. The type of feedback that participants received influenced their average confidence, and the feedback also influenced the extent to which participants' confidence changed over time (i.e., with task experience). If the error in the confidence elicitation process is truly unsystematic, then error magnitudes should not change across these experimental conditions. Thus, something else must be driving the observed changes in confidence.

The extent to which true experimental effects can be teased from error is a pervasive question which has inspired work in statistics and statistical modeling for decades. Juslin et al. (2000) presented an analysis that, they claimed, did just that for confidence judgments. However, just as error can obscure bias, so can Juslin et al.'s procedure. Any correction, regardless of its theoretical underpinnings, runs the risk of correcting away the very effect of interest. Therefore, as we demonstrated, Juslin et al.'s correction does not provide support for the idea that overconfidence biases are smaller than they seem (or nonexistent).

Misrepresenting Juslin et al.'s position. Olsson et al. (2008) state that their intent in 2000 was to demonstrate that random error could account for typical overconfidence effects, not to show that random error is the sole cause of overconfidence or to invalidate other explanations (manuscript p. 4). Furthermore, we agree that they did show, quite convincingly, that random error can account for overconfidence effects, just as Erev et al. (1994) did. If they wish to claim that this was their only intention, we are not in a position to argue otherwise.

However, it is hardly surprising that we read more into their paper (entitled "Naive Empiricism and Dogmatism in Confidence Research") than they now insist they intended. From the abstract, where they state, "…Contrary to widespread belief, there is …little support for a cognitive-processing bias in [2-alternative general knowledge questions]" (p. 384) to their closing statement, "…We propose that there is little reason to further bolster the hypothesis of positive biases by pointing to a cognitive overconfidence bias in the processing of general knowledge" (p. 394), it seemed to us that they were arguing that

overconfidence and the hard-easy effect were almost entirely, if not completely, artifactual. If these statements were supposed to be interpreted to mean only that random error is an important component of overconfidence, we beg the authors' pardon.

We were not the only readers of this paper to have overinterpreted the authors' position. Their paper is held up repeatedly by other researchers as supporting the view that overconfidence is artifactual. For example, Gigerenzer (2004) states that while we once believed that overconfidence and the hard-easy effect were due to bias, now, thanks to Juslin et al. (2000), we know the effects instead are due to "environmental effects" (tricky questions and random error) on unbiased people (see Gigerenzer's Table 1). Gu and Wallsten (2001) cite Juslin et al. as supporting the position that overconfidence is nothing but a statistical artifact (p. 552). The authors of Juslin et al. also provide the following quotes in other papers of their own:

> Other researchers have pointed out that the overconfidence often found in empirical studies may just be a product of measurement errors and sample biases rather than a real cognitive phenomenon (see Juslin et al., 2000, for a discussion). (Jönsson, Olsson, & Olsson, 2003, p. 31)

> With the half-range and the full-range formats, there are few signs of an underlying cognitive processing bias once one controls for the statistical effects and biases in the task selection (Juslin, Winman, & Olsson, 2000). (Winman, Hansson, & Juslin, 2004, p. 1167)

Even if all of us have misrepresented the point of Juslin et al.'s paper, it seems that, at the very least, "Naive Empiricism and Dogmatism in Confidence Research" trivializes the idea that systematic overconfidence biases exist.

It is our opinion, based on our analysis, that random error is an insufficient explanation for overconfidence. It is also our opinion that any analysis that attempts to

pull apart random error from systematic biases is fraught with difficulty. At this point, research needs to address where error comes from, which takes us back to thinking about the cognitive processes that give rise to confidence. This still-unresolved problem should encourage more studies on overconfidence, especially those incorporating experimental manipulations that can tease out the effects of different kinds of error. Precise mathematical models of the cognitive processes involved can also help us to understand the interactions between systematic and random components on the process.

References

Budescu, D. V., Wallsten, T. S., & Au, W. T. (1997). On the importance of random error in the study of probability judgment. Part II: Applying the stochastic judgment model to detect systematic trends. *Journal of Behavioral Decision Making*, *10*, 173-188.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*, 519-527.

Fischer, I., & Budescu, D. V. (2005). When do those who know more also know more about how much they know? The development of confidence and performance in categorical decision tasks. *Organizational Behavior and Human Decision Processes*, *98*, 39-53.

Gigerenzer, G. (2004). The irrationality paradox. *Behavioral and Brain Sciences*, *27*, 336-338.

Gu, H., & Wallsten, T. S. (2001). On setting response criteria for calibrated subjective probability estimates. *Journal of Mathematical Psychology*, *45*, 551-563.

Jönsson, F. U., Olsson, H., & Olsson, M. J. (2003). Odor emotionality affects the confidence in odor naming. *Chemical Senses*, *30*, 29-35.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, *107*, 384-396.

Juslin, P., Winman, A., & Olsson, H. (2003). Calibration, additivity, and source independence of probability judgments in general knowledge and sensory discrimination tasks. *Organizational Behavior and Human Decision Processes*, *92*, 34-51.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? The calibration of probability judgments. *Organizational Behavior and Human Performance*, *20*, 159-183.

Merkle, E. C., & Van Zandt, T. (2006). An application of the Poisson race model to confidence calibration. *Journal of Experimental Psychology: General*, *135*, 391-408.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin and Review*, *12*, 573-604.

Sieck, W. R., Merkle, E. C., & Van Zandt, T. (2007). Option fixation: A cognitive contributor to overconfidence. *Organizational Behavior and Human Decision Processes*, *103*, 68-83.

Sieck, W. R., & Yates, J. F. (2001). Overconfidence effects in category learning: A comparison of connectionist and exemplar memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1003-1021.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, *67*, 201-221.

Wallsten, T. S. (1996). An analysis of judgment research analyses. *Organizational Behavior and Human Decision Processes*, *65*, 220-226.

Windschitl, P. D., & Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 198-215.

Winman, A., Hansson, P., & Juslin, P. (2004). Subjective probability intervals: How to reduce overconfidence by interval evaluation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1167-1175.

Footnotes

[1]The "trial-by-trial" terminology was previously used by Budescu, Wallsten, & Au (1997).